DOI: 10.22616/RRD.30.2024.013

CHEMOMETRICS AS AN AID TO QUICKLY EVALUATE GALACTOMANNANS THROUGH INFRARED SPECTROSCOPY

***Octavio Calvo-Gomez, Akbarali Ruzibayev, Shakhnozakhon Salijonova, Shakhnozakhon Gaipova, Sarvar Khodjaev, Zulfiyakhon Khakimova, Dilshod Rakhimov**
Tashkent Institute of Chemical Technology, Uzbekistan
*Corresponding author's e-mail: solinsabajio@yahoo.com

Abstract

Galactomannans, composed of galactose and mannose, may form gels and are considered safe because of their non-toxic, biodegradable, and biocompatible nature. As a result, they are widely utilized in the food industry as stabilizers and thickeners. Among galactomannan producing species, guar gum and locust bean gum are particularly important due to their economical relevance. Guar gum and locust bean gum are often adulterated with cellulose gums like xanthan gum and carboxymethyl cellulose (CMC). Adulteration of galactomannans with other gums may introduce uncertainties regarding functionality and complicate quality control, posing a potential problem for the food industry. Among the different techniques which have been used for determining and characterizing galactomannans, Fourier Transform Infrared Spectroscopy stands out. Especially when coupled to Attenuated Total Reflection (ATR), analyses are performed rapidly, with a minimum sample preparation, and without the need for solvent or previous extraction mechanisms. However, food is a very complex matrix that contains a high number of components which generate a multitude of spectral information and large data sets. Consequently, additional data processing tools such as chemometrics are needed to be able to draw useful information from spectra. Our goal in this work is to show how to optimize conditions for instrumental analysis by infrared spectroscopy of galactomannans and its constituent monomers and create a chemometric model where galactomannans could be differentiated as a single group. We successfully optimized the PCA model obtained after chemometric processing of infrared data through reducing dimensions by loadings selection.

Key words: ATR-FTIR, Chemometrics, Galactomannans, Crops, Food Supply.

Introduction

Galactomannans are polysaccharides made of galactose and mannose in a structure with a backbone of (1->4)-linked β-D-mannopyranosyl units with side chains of (1->6)-linked α-D-galactopyranosyl units. They form viscous solutions in aqueous media (gels), thus having been extensively used in food industry mainly as thickeners and stabilizers. Galactomannans naturally occur in various plant seeds (mostly from leguminosae family), although they may occur in fungal species as well (Srivastava & Kapoor, 2005). Among galactomannan producing species, guar gum (from *Cyamopsis tetragonoloba*) and locust bean gum (from *Ceratonia siliqua*, also called carob) are particularly important due to their economical relevance (Prado *et al.*, 2005; Dakia *et al.*, 2008). Although galactomannans may have varying physicochemical properties depending on the galactose: mannose ratio, they are in general regarded as safe due to nontoxicity, biodegradability, and biocompatibility (Sharma, Kumar, & Sharma, 2020).

Guar gum and locust bean gum are often adulterated with cellulose gums like xanthan gum and carboxymethyl cellulose (CMC) in the food industry, which are generally cheaper thickeners and stabilizers (Prado *et al.*, 2005). However, this adulteration can cause problems. Blends of these gums were found to have varying effects on the flow properties of emulsions, with interactions between the different polysaccharides affecting stability (Nor Hayati, Wai Ching, & Rozaini, 2016). Furthermore, molecular interactions between xanthan gums and galactomannans like guar gum can complicate characterization of such mixtures (Schreiber *et al.*, 2020). Therefore, adulteration of galactomannans with other gums may introduce uncertainties regarding functionality and complicate quality control, posing a potential problem for the food industry (Flurer, 2000). Various analytical methods have been used to detect adulteration in galactomannans. Some of them are specific for determining galactomannans/ cellulose gels/ mixed systems' properties, like the one developed by Fernandes, where periodate oxidation is used (Fernandes, 1994). And some others are based in widely used techniques such as infrared spectroscopy (Prado *et al.*, 2005). In general, mid-infrared spectroscopy (IR) is a rapid and simple technique which has been shown to be a valuable tool for determining adulteration and authenticity of various foods, including galactomannans. Mendes and Duarte identified intense absorption bands in the region between 950 and 700 cm^{-1} which were able to correlate to the presence of adulterants such as starch, different to the polysaccharides which normally exist in coffee, where galactomanans are included (Mendes & Duarte, 2021). Prado *et al.* used Fourier Transform Infrared Spectroscopy (FTIR) to differentiate among different type of carbohydrate gums and mixtures, including galactomannans (Prado *et al.*, 2005).

In Fourier Transform Infrared Spectroscopy (FTIR), molecules absorb infrared radiation due to changes in the dipole moment of chemical bonds, thus the wavelengths of the absorbed light will depend on the structure of their functional groups; in this manner, individual bands may be linked to specific functional groups. Therefore, structural information from the molecules is gathered and displayed in the infrared spectrum of a compound or mixture of compounds. An important advantage of FTIR, especially when coupled to Attenuated Total Reflection (ATR), is that many different compounds may be analyzed including liquids, powders, polymers, or semisolids. Besides, those analyses are performed rapidly, with a minimum sample preparation, and without the need for solvent or previous extraction mechanisms (Smith, 2018; Tiernan, Byrne, & Kazarian, 2020).

However, food is a very complex matrix that contains a high number of components which give rise to a multitude of spectral information and large data sets. Consequently, fast statistic and mathematical analyses are needed to fully understand all the complexity of data, as well as to be able to draw useful information from it (Roberts & Cozzolino, 2016). Therefore, chemometrics, which may be defined as 'the chemical discipline that uses mathematical and statistical methods, (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum chemical information by analyzing chemical data' is an invaluable tool that allows for gathering information which is not normally visible in a group of spectra of different compounds, where contribution from many different functional groups from many compounds may cause overlapping to occur (Otto, 2016).

Infrared spectra are obtained after sampling radiation absorption through wavelengths that correspond to the mid-infrared section of the electromagnetic spectrum (4000-400 cm^{-1} approximately). Consequently, there are several variables because each sampled wavelength constitutes one. Accordingly, in a spectrum with 2300 wavelengths, there are 2300 variables. For optimizing a chemometric model for IR spectra, variable selection is important for reducing dimensionality and complexity, thus improving model performance and interpretability. Some wavelength regions may contain mostly noise with little chemical information, hence removing these regions improves the signal-to-noise ratio. By selecting characteristic wavelengths or wavelength intervals, the interpretability of the model can be strengthened. Moreover, the wavelengths selected provide insight into the molecular or atomic transitions that are most influential for a given analytical problem. This may aid in chemical interpretation (Yun, 2022). Advantages of reducing dimensionality include reduced risk of overfitting, better model interpretability, and reduced computational cost and time (Lee, Liong, & Jemain, 2018).

A supervised algorithm in chemometrics is a classification or regression method that learns from example inputs in a training dataset that contain labels in order to predict the target labels of new unseen instances. Common supervised algorithms include partial least squares regression (PLS) and linear discriminant analysis (LDA), which have been used extensively in quantitative and qualitative analysis of spectroscopic and chromatographic data. In contrast, an unsupervised algorithm in chemometrics is an exploratory technique that groups or segments a dataset without using labels in order to discover hidden patterns in the data. Examples of unsupervised algorithms are principal component analysis (PCA), cluster analysis, and self-organizing maps (SOM), which have found applications in areas such as process monitoring and fingerprinting to detect outliers or identify new classes. Unsupervised algorithms are generally used for pattern recognition and dimension reduction without prior knowledge of the desired outputs (Wold, Sjöström, & Eriksson, 2001; Geladi, 2003).

PCA is a non-supervised algorithm which may be used to reduce the dimensionality of large data sets by transforming a number of correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Loadings represent the correlation between each original variable and the components and can be used to interpret the underlying structure of the components. The loadings identify which original variables contribute most strongly to each component. Variables with high loadings, either positive or negative, on a component are the ones most represented by that component. Therefore, the loadings aid in dimensionality reduction in PCA by identifying which original variables have the strongest influence on the principal components and which variables can potentially be excluded from further analysis without much loss of information (Abdi & Williams, 2010; Jolliffe & Cadima, 2016).

Our goal in this work is to show how to optimize conditions for instrumental analysis by infrared spectroscopy of galactomannans and its constituent monomers and create a chemometric model where galactomannans could be differentiated as a single group. Although papers previously published have reported algorithms for discriminating between different types of galactomannans after FTIR-chemometric analyses (Prado *et al*., 2005), we are aiming to develop a quick and simple method for discerning galactomannans as a group, while using their individual monomers, mannose and galactose, as reference materials for establishing the model. We also wanted to use an unsupervised algorithm for doing so, to eliminate the requirement of preassigning classifications to specific categories during data processing. Optimization of the model was done through dimensionality reduction following purely chemometric criteria such as loadings associated to wavelengths. We do not claim that this work would be a stand-alone test but rather a supportive material for helping in the development of a tool for a quick identification of adulterants in galactomannans used as food additives.

Materials and Methods

Materials: 8 samples of guar gum (from *Cyamopsis tetragonoloba*) (identified in this work as GUA) and 7 samples of carob -locust bean gum- (from *Ceratonia siliqua)* (identified in this work as ALG -for 'algarrobo', carob in Spanish-) were purchased from local producers. Mannose (MAN) and galactose (GAL) standards were sourced from Sigma-Aldrich (USA).

Equipment: Agilent Cary 660 Fourier Transform Infrared Spectrophotometer (Agilent, USA) equipped with a Pike Technologies germanium crystal ATR (Pike Technologies, USA).

Software: Resolutions Pro (Agilent, USA), Spectragryph (Friedrich Menges Software-Entwicklung, Germany), MS Excel from Microsoft 365 (Microsoft, USA), Pirouette (Infometrix, USA) and JMP Pro (SAS Institute, USA).

Methods: Both guar gum and carob are powders thus the process for instrumental analysis had to be optimized. Since ATR works by sending an evanescent wave into the sample, only that infinitesimal part of the sample that comes into contact with the crystal, specifically within the distance where the evanescent wave penetrates into the sample, will actually provide information regarding absorption of infrared light by the different functional groups found there. Hence, in the infrared spectrum will only be information of that part of the sample, whose specific depth will depend on the crystal of the ATR (because evanescent waves depend on the type of crystal) (Smith, 2018; Tiernan, Byrne, & Kazarian, 2020). Beyond it, any additional sample will be neglected, therefore, a specific weight of sample is not the appropriate way for guaranteeing a proper data acquisition during instrumental analysis.

It is important to guarantee that the part of sample in contact with the evanescent wave is optimized. Thus, before analysis, several parameters were optimized. Since samples are solid powders, different laboratory spoons were used for optimizing sample size. The use of the included clamp accessory of the ATR was evaluated, and finally, the amount of time after placing the sample in the crystal was also considered (since samples may be hygroscopic). Also, cleaning protocols were assessed.

Our final conditions were, regarding the amount of sample, to take about 20 mg of sample (or standard), place them on top of the ATR crystal, and then, using a cardboard aid (not a metallic spatula to prevent damaging the crystal), accommodate it while ensuring that no part of it remain uncovered. We realized that, considering the hygroscopicity of some of the samples, the time elapsed since their placement in the crystal and the time when the readings were performed could bring some changes in the spectra. We also realized that pressing the sample with the clamp accessory of the ATR improved the consistency of obtained spectra. Therefore, we concluded that readings should be taken immediately after securing the swivel pressure tower of the clamp accessory. Additional conditions included the removal of dust and cleaning of both the crystal and plate of the ATR with isopropyl alcohol, while allowing one minute afterwards in order to allow for alcohol evaporation, and to take background reading between samples. Each sample was read 32 times (32 scans) from 750 to 3700 cm^{-1} and a wavenumber distance of 4 cm^{-1} and resulting spectra provided the data to be used for the assembly of the sample matrices to be processed by the chemometric algorithms that followed.

Results and Discussion

In this work, we performed ATR-FTIR analyses of several samples of two different types of galactomannans, locust beam gum (carob) and guar gum, as well as of their constituents, monosaccharides mannose and galactose (8 times each). Spectra of all the analyzed samples and standards are displayed in ‘Figure 1’. It is worth noting that color coding in the Figure 1 were used for all spectra of the corresponding type of either sample or standard, thus within MAN there are spectra of 8 analyses, as well as in the case of GAL and GUA. In ALG, spectra are of 7 analyses.

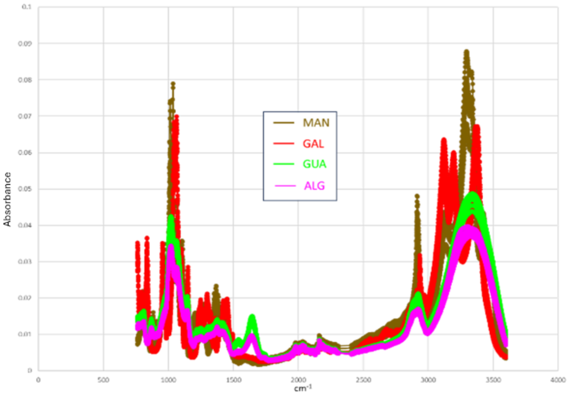

Figure 1. Raw infrared spectra of all the samples analyzed in this work (in absorbance).

Since we analyzed both galactomannans and galactose and mannose standards, one conceivable approach to the identification of galactomannans from possible adulterants was to focus on absorption bands present in both samples and standards. However, infrared spectra may have contributions from many sources including fundamental bands, overtones, and possible presence of groups with overlapping absorption bands; therefore, simple interpretation of infrared spectra from complex mixtures may become difficult, as depicted in ‘Figure 1’. Therefore, we decided to conduct chemometric analyses to differentiate and analyze the data effectively. By utilizing these analytical techniques, we can uncover meaningful insights and obtain useful information.

Raw spectra contained 2309 wavelengths (thus variables). Although explored, no spectra pre-processing (such as Savitsky−Golay smoothing, first or second derivative, normalization, or rubber band correction) other that background subtraction was performed on the spectra. Therefore, raw absorbance data was the information considered for chemometrics in this study. After assembling the data matrix, dimensions were reduced considering not functional groups nor regions but merely chemometric criteria, in this case loadings.

As explained in introduction, supervised algorithms require pre-assignment to a given category. If we are developing a method aiming for a further detection of adulterants, we do not want to bias the model by assigning any beforehand label. For this reason, an unsupervised algorithm was used to determine dispersion, in this case, Principal Component Analysis (PCA). 'Figure 2' shows the 2D PCA of the complete spectra, with all the wavelengths (a total of 2309) obtained by analysis in the ATR-FTIR equipment.

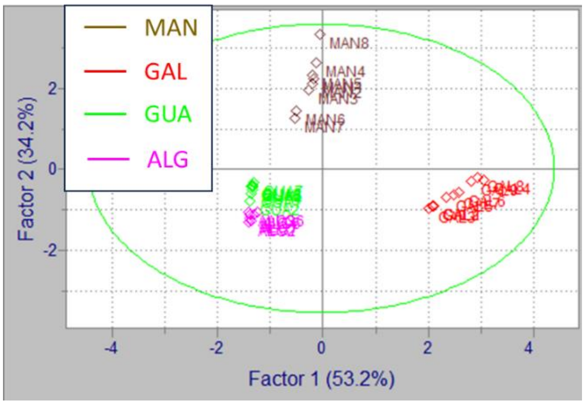

Figure 2. PCA of raw infrared spectra of all analyzed samples and standards (2309 variables).

Although clear groupings between the samples are noticeable at first glance, there is also a certain dispersion among the different repetitions in both standards, especially in the case of MAN. This dispersion could be related to moisture absorption, as during the optimization process, visible changes in the spectra were observed over time after being placed onto the ATR platform. However, this dispersion only affected some wavelengths, which is why dimension reduction was sought to decrease the dispersion among the different repetitions of both standards (and thus eliminate possible interferences from elements not part of the standards). Besides, GUA and ALG, although very close together, still may be distinguished as two different groups.

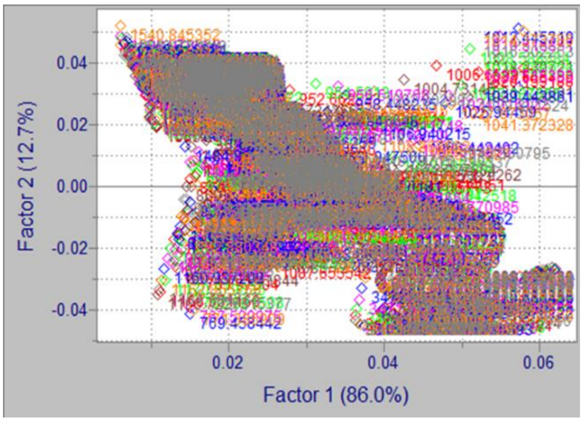

Figure 3. Loadings of PCA of raw infrared spectra of all analyzed samples and standards (2309 variables).

For improving the model, the analysis of the 'loadings' was carried out. Loadings reflect the contribution that each specific variable (in this case, each wavelength) makes to the system's dispersion, as manifested in the absolute value they have for each principal component. The higher this value, meaning further from zero, the greater the contribution of that variable to the system's dispersion. In 'Figure 3', the graphical representation of all the loadings in the PCA model depicted in 'Figure 2' is displayed. The 2309 variables are included in 'Figure 3'. Those farther away from the center account for the higher dispersion, while those in the very center bring the least contribution to dispersion. Thus, the first step to dimension reduction involves identifying and gradually removing variables with the lowest contributions according to the loadings. 'Figure 4' is the graphic representation of loadings after said modification.

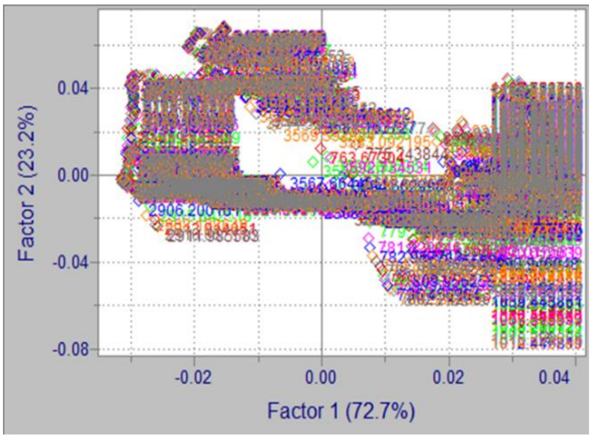

Figure 4. Loadings of PCA of raw infrared spectra of all analyzed samples and standards after removing wavelengths with minimum contribution to dispersion.

For further optimization of the model, we used supervised algorithms such as K-Nearest Neighbor (KNN), Soft independent modelling of class analogies (SIMCA), Alternate least squares (ALS), and Partial least squares – Discriminant analysis (PLS-DA) in order to identify those wavelengths that account for maximum separation into categories. However, it is important to note that supervised algorithms were used not as the final model but as an aid for wavelength selection for improving the PCA model. In 'Figure 5', the KNN of the spectra is analyzed.

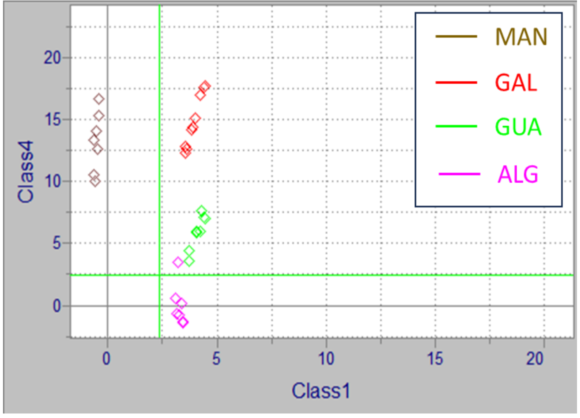

Figure 5. KNN of raw infrared spectra of all analyzed samples and standards.

Ultimately, a new matrix remains where separation between members of the same set will be minimized, while separation between different sets will be maximized. 'Figure 6' represents the PCA created with this matrix, now with only 888 variables.

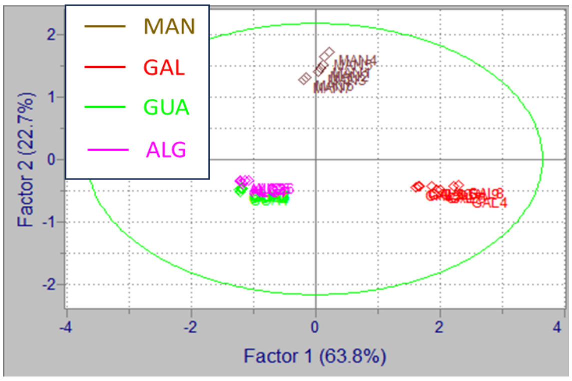

Figure 6. PCA of dimensionality-reduced infrared spectra of all analyzed samples and standards (888 variables).

Comparing 'Figure 2' with 'Figure 6', it is noticeable how the groups of both MAN and GAL 'compressed' after the variable reduction by 'distillation' of loadings. Likewise, the groups of both GUA and ALG also compressed and even overlapped, thus now constituting a single set, as originally intended.

Conclusions

1. ATR-FTIR followed by chemometrics is an excellent technique for analyzing galactomannans that are used as stabilizers and thickeners in the food industry.
2. Without chemometrics, useful information which may be obtained from infrared spectra of samples is limited.
3. An unsupervised chemometric model, PCA, was successfully optimized through reducing dimensions by loadings selection.

Acknowledgements

Authors thank Dr. Mercedes G. Lopez from Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (Cinvestav-IPN) Campus Guanajuato, Irapuato, Mexico, for kind help and assistance.

References

Abdi, H. & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*(4), 433-459. DOI: 10.1002/wics.101.

Dakia, P. A., Blecker, C., Robert, C., Wathelet, B., & Paquot, M. (2008). Composition and physicochemical properties of locust bean gum extracted from whole seeds by acid or water dehulling pre-treatment. *Food Hydrocolloids*, *22*(5), 807-818. DOI: 10.1016/j.foodhyd.2007.03.007.

Fernandes, P. B. (1994). Determination of the physical functionality of galactomannans in xanthan gum/galactomannan mixed systems by periodate oxidation. *Food Control*, *5*(4), 244-248. DOI: 10.1016/0956-7135(94)90024-8.

Flurer, C. L. (2000). Characterization of galactomannans by capillary electrophoresis. *Food Additives and Contaminants*, *17*(9), 721-731. DOI: 10.1080/026520300415255.

Geladi, P. (2003). Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochimica Acta Part B: Atomic Spectroscopy*, *58*(5), 767-782. DOI: 10.1016/S0584-8547(03)00037-5.

Jolliffe, I. T. & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. DOI: 10.1098/rsta.2015.0202.

Lee, L. C., Liong, C.-Y., & Jemain, A. A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. *The Analyst*, *143*(15), 3526-3539. DOI: 10.1039/C8AN00599K.

Mendes, E. & Duarte, N. (2021). Mid-Infrared Spectroscopy as a Valuable Tool to Tackle Food Analysis: A Literature Review on Coffee, Dairies, Honey, Olive Oil and Wine. *Foods*, *10*(2), 477. DOI: 10.3390/foods10020477.

Nor Hayati, I., Wai Ching, C., & Rozaini, M.Z.H. (2016). Flow properties of o/w emulsions as affected by xanthan gum, guar gum and carboxymethyl cellulose interactions studied by a mixture regression modelling. *Food Hydrocolloids*, *53*, 199-208. DOI: 10.1016/j.foodhyd.2015.04.032.

Otto, M. (2016). *Chemometrics: Statistics and Computer Application in Analytical Chemistry* (1.^{a} ed.). Wiley. DOI: 10.1002/9783527699377.

Prado, B. M., Kim, S., Özen, B. F., & Mauer, L. J. (2005). Differentiation of Carbohydrate Gums and Mixtures Using Fourier Transform Infrared Spectroscopy and Chemometrics. *Journal of Agricultural and Food Chemistry*, *53*(8), 2823-2829. DOI: 10.1021/jf0485537.

Roberts, J. J. & Cozzolino, D. (2016). An Overview on the Application of Chemometrics in Food Science and Technology—An Approach to Quantitative Data Analysis. *Food Analytical Methods*, *9*(12), 3258-3267. DOI: 10.1007/s12161-016-0574-7.

Schreiber, C., Ghebremedhin, M., Zielbauer, B., Dietz, N., & Vilgis, T. A. (2020). Interaction of xanthan gums with galacto- and glucomannans. part I: Molecular interactions and synergism in cold gelled systems. *Journal* *of Physics: Materials*, *3*(3), 034013. DOI: 10.1088/2515-7639/ab9ac8.

Sharma, D., Kumar, V., & Sharma, P. (2020). Application, Synthesis, and Characterization of Cationic Galactomannan from Ruderal Species as a Wet Strength Additive and Flocculating Agent. *ACS Omega*, *5*(39), 25240-25252. DOI: 10.1021/acsomega.0c03408.

Smith, B. (2018). *Infrared Spectral Interpretation: A Systematic Approach* (1.ª ed.). CRC Press. DOI: 10.1201/9780203750841.

Srivastava, M. & Kapoor, V.P. (2005). Seed Galactomannans: An Overview. *Chemistry & Biodiversity*, *2*(3), 295-317. DOI: 10.1002/cbdv.200590013.

Tiernan, H., Byrne, B., & Kazarian, S.G. (2020). ATR-FTIR spectroscopy and spectroscopic imaging for the analysis of biopharmaceuticals. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *241*, 118636. DOI: 10.1016/j.saa.2020.118636.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109-130. DOI: 10.1016/S0169-7439(01)00155-1.

Yun, Y. H. (2022). Wavelength Selection Methods. En X. Chu, Y. Huang, Y.-H. Yun, & X. Bian, *Chemometric Methods in Analytical Spectroscopy Technology* (pp. 169-207). Springer Nature Singapore. DOI: 10.1007/978-981-19-1625-0_5.